Visual Analytics (INM433)

Daniel Dixey

MSc Data Science
Wednesday 8th April

Data

Overview, description and properties of the dataset

Chosen Dataset
- Competition Data from the Crossfit Open (2011)
- Worldwide competition
- 5 Week long competition, each week represents one demanding workout (WOD)

Contents of the dataset
- Basic Athlete metrics: Weight, Age, Gender and Height
- Results and Rankings

Dataset Source
- Acquired by Web Crawling
- Expectation: Messy and unprocessed

What is Crossfit?
- CrossFit is a rigorous fitness methodology that that attempts to unite multiple domains of fitness into one programme
- Domains include; Weightlifting, High Intensity Interval Training (HIIT) and Gymnastics
- Originated from the San Francisco/ Bay Area in 2004
- Approximately 10000+ Boxes Worldwide

Data [10%]. Description of the data chosen for the analysis: type, structure, size, properties of the components.

Data

Overview, description and properties of the dataset

Type of Data: Multi-dimensional dataset
Class of Data: Object Referenced Data

Data Quality:

Incompleteness
- For all the attributes there exists a degree of incompleteness
- Consideration will be determined on a feature by feature basis

Uncertainty
- The credibility and integrity of the data could be of some concern,
- Where necessary spurious and obviously false (age >100 years and weight <20kg) information and will be removed where deemed necessary during the analysis

Number of Instances	25973		
Number of Referrers	2	Athletes Name	
		Competition Region	
Number of Attributes	19	Nominal	2
		Ordinal	3
		Interval	-
		Ratio	14
Types of Values		Numeric	18
		Textual	3
		Spatial	0
		Addresses	0
		Temporal	1

Notes: The dataset overview above is the dataset has not been cleaned prior to review. 'athlete id' and 'nameURL' have been considered as one value. The region feature can be considered as both textual and temporal.

Data [10%]. Description of the data chosen for the analysis: type, structure, size, properties of the components.

Analysis Tasks

Literature review, work to be undertaken and synoptic task

Literature Review: Dimensionality Reduction (DR)

Exploitation: The focus of the review was deliberately limited to DR as the number of considerations and techniques available in the area of multidimensional analysis are vast.

Limitation: DR was specifically identified as very useful in assisting with the visualisation of multidimensional datasets.

Challenges:
- Trying to understand the data model would find it difficult to interpret and make meaningful interpretations of the data
- Minimising Information loss

Work Accessed and Discussed:
- Visual Hierarchical Dimensionality Reduction (VHDR) – Methodology for avoiding gaining maximum information
- Neural Networks (Autoencoders) – An alternative approach to DR, can "learn" non-linearities
- Igloo Plots – A visual approach to visualising many dimensions simultaneously
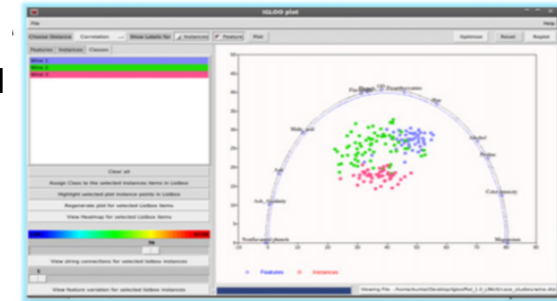
Analysis tasks [10%]. Analysis task(s) chosen for your analysis: Do these relate to the tasks you addressed in your literature review (part 1)? If so, how? Specific task formulations for the chosen data and the corresponding generic task types.

Analysis Tasks

Literature review, work to be undertaken and synoptic task

Research Question:

"Compare the variations in athlete scores to identify cohorts of high performance"

Intermediate Questions

- Who are the top individuals globally and by region? Are they comparative and exhibit the same types of characteristics?
- How does the drop of participation of each of the Open events distort the comparison of athletes?
- What key features do the top athletes have that the less competitive do not?
- How do gender, height and weight relate to the performance of athletes in the Crossfit Open competition?

It is expected that during the analysis that further questions (elementary and intermediate) will arise as a result of these pre-determined questions.

Analysis tasks [10%]. Analysis task(s) chosen for your analysis: Do these relate to the tasks you addressed in your literature review (part 1)? If so, how? Specific task formulations for the chosen data and the corresponding generic task types.

Analysis Methodology

Overview, description and techniques used in the process

5 Part Procedure

1. **Understanding of the characteristics of the attributes**
   - Human Labour - *understanding of each of the distributions and basics statistics about every feature*

2. **Understanding of the Relationships between the attributes**
   - Human Labour – *Investigate the relationships between features perception*
   - Computer Labour - *Using a variety of different visual display types*
   - Computer Labour - *Careful consideration of the types of visual encodings: shape, size, colour and labelling, to aid and complement with the understanding*

3. **Describing the relationships**
   - Computer Labour - *Two computational methods will be utilised at this stage; Gaussian Mixture Models (GMM) and Self Organising Maps (SOM)*
   - Human Labour – *Answer the intermediate and research questions*
   - Computer Labour – *Does the use of an Autoencoder provide significantly better insight and knowledge?*

4. **Exploring the findings iteratively through interaction**
   - Human and Computer Labour - *It is expected that navigating many times between phase 1-3*

5. **Final presentation**
   - Human and Computer Labour - *Final Improvement and refinement of the visualisations*

Analysis methodology [30%]. Methodology of your analysis.

Implementation

Software and process work flow

Preprocessing   Exploratory Analysis   Final Presentation

LibreOffice
Calc

Python
Orange Canvas

Python
Seaborn
"attractive statistical graphics"

Javascript
D3.js
"helps you bring data to life"

- *Interactive work flows*
- *Widgets*
- *Machine learning*
- *Transformation*
- *Testing and evaluation*

Implementation [30%]. Implementation of the analysis methodology: software used, links between methods (integrated in the same software or data transfer)

Implementation

- Software and process work flow

Screen shot of *Orange Canvas*

Implementation [30%]. Implementation of the analysis methodology: software used, links between methods (integrated in the same software or data transfer)

CITY UNIVERSITY LONDON

8

Analysis process

Work to be undertaken and results

Display initial results

Recite all of the lecture notes here

Analysis process [10%]. Illustrated description of the analysis process

Results and conclusion

Current status of the report/analysis

Work is still ongoing...

Report and analysis progress: *70%*

Results and conclusion [10%]. To what extent the posed task(s) have been fulfilled?